View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Sussex Research Online

Mutational patterns in oncogenes and tumour suppressors

Article (Accepted Version)

Baeissa, Hanadi M, Benstead-Hume, Graeme, Richardson, Christopher J and Pearl, Frances M G (2016) Mutational patterns in oncogenes and tumour suppressors. Biochemical Society Transactions, 44 (3). pp. 952-931. ISSN 0300-5127

This version is available from Sussex Research Online: http://sro.sussex.ac.uk/60335/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk

Mutational patterns in oncogenes and tumour suppressors

Hanadi M Baeissa[1], Graeme Benstead-Hume[1], Christopher J Richardson[2] and Frances M G Pearl[1*]

1: Bioinformatics Group, School of Life Sciences, University of Sussex, Falmer, Brighton

2: Division of Structural Biology, The Institute of Cancer Research, London

* Corresponding author
email: f.pearl@sussex.ac.uk

Abstract

All cancers depend upon mutations in critical genes, which confer a selective advantage to the tumour cell. Knowledge of these mutations is crucial to understanding the biology of cancer initiation and progression, and to the development of targeted therapeutic strategies. The key to understanding the contribution of a disease-associated mutation to the development and progression of cancer, comes from an understanding of the consequences of that mutation on the function of the affected protein, and the impact on the pathways in which that protein is involved.

In this paper we examine the mutation patterns observed in oncogenes and tumour suppressors, and discuss different approaches that have been developed to identify driver mutations within cancers that contribute to the disease progress. We also discuss the MOKCa database where we have developed an automatic pipeline that structurally and functionally annotates all proteins from the human proteome that are mutated in cancer.

Introduction

In most diseases of genetic origin, the disease phenotype can usually be attributed to a small number of defined mutations, which once located are readily distinguished from the essentially wild-type genetic background [1]. Cancer is also fundamentally a genetic disease, with the phenotype arising by somatic acquisition of a set of defined 'hallmark' mutations [2]. These exert their effect by activating oncogenes and/or inactivating tumour suppressors, one or more of which may already be mutated in the germline in inherited cancer predispositions syndromes.

Acquisition of the genetic changes that confer hallmark traits of invasive cancer depends on loss of genetic stability early in the tumour cell lineage typically initiated by a defect in the DNA damage response (DDR) [3]. Paradoxically, the inherent genetic instability that gives tumours their evolutionary plasticity underlies their sensitivity to the genotoxic drugs and radiation that constitute many first-line cancer therapies. An important consequence of this genetic instability is the presence of large numbers of mutational changes in the genomes of tumours as compared to untransformed cells from the same individual [4]. The overwhelming majority of these changes may be inconsequential in terms of driving the cancer phenotype, but generate a high level of mutational 'noise' within which the significant driving mutations may be very difficult to identify.

There has been a substantial increase in understanding of the many pathways that can drive the hallmark traits of cancer in the last few years[5], and many specific inhibitors of the proteins that constitute those pathways have been developed. Together with the development of rapid and low-cost genome sequencing, there is now the real prospect of 'personalised' drug therapies precisely targeted to the idiosyncratic regulatory malfunctions resulting from the mutations that drive an individual cancer [6], so long as these can be distinguished from the substantial background of irrelevant 'passenger' mutations, so that the genotype can be used to predict the phenotype .

Given the large numbers of mutations typically observed [7] experimental determinations of the consequences on protein function of the individual

mutations observed in a cancer genome are not realistic, and computational approaches are required.

Identifying Driver Genes

These are several statistical approaches (eg [8, 9]) that identify significantly mutated genes within large cohorts of sequenced tumours. These approaches are very good at identifying highly recurrent mutated genes but as yet, the data sets are not large enough to have the statistical power to detect low frequency mutated genes that contribute to the initiation and progression of cancer. This can pose a problem because although a few genes are highly mutated, the majority of somatic mutations occur in genes that are infrequently mutated [10, 11].

Characteristics of Tumour Suppressors and Oncogenes

Driver genes are classified by the manner in which, when mutated, they contribute to the disease process. Tumour suppressors contribute to the development of cancer when mutations (or in some instances epigenetic silencing) result in their loss of function (LOF). The alterations to these genes are generally molecularly recessive where both copies of the gene require a LOF defect to cause disease [12]. For instance, this may be a truncation or missense mutation on one allele, combined with a complete loss of the second. This commonly occurs in kidney renal clear cell carcinoma (KIRC), where the loss of the chromosome arm 3p in KIRC combined with concurrent mutations on the remaining allele results in complete ablation of functioning VHL [13].

In oncogenes, an increase in activity, or a change of function is required for tumorigenesis. They tend to exhibit a molecularly dominant mode of action, and usually only one defective copy of the gene is required to provide an oncogenic phenotype. This is exhibited in BRAF where V600E activating mutations constitutively activates BRAF in malignant melanoma [14], or in BCR-ABL in chronic myelogenous leukaemia where a translocation constitutively activates ABL-kinase.

Missense mutations in tumour suppressors can result in its loss of function in a variety of manners including loss of stability of the protein or the disruption of a crucial ligand/DNA/protein-binding site [15]. In cohorts of tumours, these mutations are often liberally dispersed along the length of the gene, as protein function can be disrupted by mutations at a multitude of positions [16]. Conversely, in oncogenes, driver missense mutations tend to cluster at distinct locations in the amino acid sequence impacting on sites of protein-protein interaction, allosteric regulation, post-translational modification or ligand-binding. Often only a very few, specific mutations can lead to activation of the protein product or a change of a protein function [16].

Identifying driver mutations

Sequence and structural data have been utilised to predict whether a missense mutation or a small insertion or deletion could be disease-causing using a variety of approaches. Sequence conservation is used to predict which mutations can be tolerated within a protein structure, and similarly, protein structures have been used for estimating how disruptive a missense

mutation may be [15, 17-20]. Techniques originally developed to predict the consequences of amino-acid changes observed in SNPs and Mendelian genetic diseases, have been applied to cancer mutations, but have often failed to provide sufficiently reliable prediction.

More recently algorithms have been specifically developed to distinguish cancer-associated somatic driver mutations from passenger mutations. These include profile-based methods for assessing missense mutations (eg [21-24]), and machine learning algorithms for assessing the pathogenicity of missense mutations [25] and indels [26].

Approaches to distinguish between tumour suppressors and oncogenes

As the mutational patterns observed in cohorts of tumour samples clearly differ between tumour suppressor and oncogenes, several groups have used this information to automatically distinguish between them. For instance, Vogelstein's 20:20 rule [16] states that if 20% of all mutations observed in a gene within a cohort of tumour samples are truncations, then that gene is likely to be a tumour suppressor, where as if 20% of all missense mutations occur at a single position in the sequence, the gene is predicted to be an oncogene. These types of patterns have also been included in machine learning algorithms to automatically distinguish between tumour suppressors and oncogenes (eg [27]) using data from whole exome sequencing.

MOKCa database

The MOKCa database [28] (http://strubiol.icr.ac.uk/extra/mokca/) was developed to structurally and functionally annotate, and where possible predict, the phenotypic consequences of disease-associated mutations in protein kinases implicated in cancer. We have recently extended the database to include all the proteins from the human genome that are mutated in cancer (see supplementary figure 1).

Somatic mutation data from the COSMIC database [7] have been mapped to their position in UniProt sequences [29]. Each mutation is described by its alteration to the protein structure, eg V600E. When a mutation has been reported on more one occasion, it is stored as an "aggregate" mutation and the number of observations of the aggregate mutation is recorded. Different genetic changes that result in the same protein coding mutation are presented together at the protein level and each disease type in which this mutation has been recorded is also presented on the protein overview page.

Functional annotations for each protein are displayed. These include the identification and position of Pfam domain assignments within the protein sequence [30], and the positions of residues effected by post-translational modifications including phosphorylation, glycosylation, and ubiquitination [31]. Gene Ontology (GO) annotations have also been obtained for each protein [32].

Structural Mapping of Mutations

The amino acid sequence for every Pfam-annotated domain for which COSMIC records a cancer-associated mutation has been scanned against the Protein Data Bank (PDB) [33] using BLAST/PSI-BLAST [34], to map the

mutation onto the protein structure of the affected human protein domains where the structure has been experimentally determined, or onto the most closely related homologous structure where the experimental structure is not known.

The positions of the individual mutations can be viewed on the mutation web page using the Jmol application [35], and the multiple sequence alignment between the query domain and the PDB template is displayed using Jalview [36].

Development of web-interface

The new web-interface for MOKCa database can accessed at http://strubiol.icr.ac.uk/extra/mokca/ (see figure 1) and can be searched by gene name or by UniProt accession [29].

Users can also browse the data using gene names either exploring the complete genome or our curated sets of genes that are implicated in cancer. These include, protein kinases, oncogenes and tumour suppressors, proteins involved in the DNA damage response (DDR) [37] and those proteins that are current targets of chemotherapy and personalised cancer medicine regimes (drug targets) [38].

Activating mutations in oncogenes

Analysis of data in the MOKCa database suggests that although there are a large number of ways to inactivate the protein product of a gene, there are probably only a limited number of ways that small mutations (missense,

truncations, indels) are able to activate them. We have identified several common mechanisms of activation - some of these are highlighted below.

Activating mutations in protein kinases

Protein kinases can be thought of being in equilibrium between the open and closed conformations. Usually, other protein kinases phosphorylate the activating residues (S/T/Y) -moving the conformational equilibrium towards the open, active conformation (See Figure 2), whereas protein phosphatases remove the phosphate groups shifting the conformational equilibrium back to the closed, inactive conformation. These processes leads to highly regulated control of the conformation and activation of kinase domains.

One of the most frequently reported mutations is the activating mutation V600E in B-Raf, a driver missense mutation in malignant melanoma. Examination of V600E mutation models using the SAAPdat tool [15] (Figure 3), clearly shows that the structural impact of the mutation differs in the active and inactive conformations of the protein. The mutation is predicted to be structurally tolerated when the BRAF kinase domain is in the open, active conformation, yet in the closed, inactive conformation the mutation is predicted to introduce a hydrophilic residue and a buried charge into the core of the protein. This would result in the destabilisation of the inactive conformation, moving the equilibrium of the protein towards the active conformation where the mutation is better tolerated.

Recent molecular dynamic simulations support this model, suggesting that the V600E mutation increases the energy barrier of the transition from the active to inactive conformation, trapping B-Raf in the active state. They also suggest

that an increase in the flexibility of the activation loop may also speed-up phosphorylation [39].

Dependant on their location within the kinase domain, missense mutations will often be better tolerated in one or other conformation of the protein kinase resulting in an alteration of the conformational equilibrium and constitutive activation (or in some cases deactivation) of the protein kinase.

Another observed mechanism for the constitutive activation of protein kinases is the loss of inhibitory phosphorylation sites. These include the auto inhibitory phosphorylation sites in KIT at position Y823 (D/C/N mutations) and the S259A mutation in the PKC phosphorylation site in Raf1, that mediates inhibitory 14-3-3 protein [40]. Tyrosine receptor kinases can also be activated by dimerization of the extracellular domains resulting in ligand-independent activation of the receptor. This is observed in FGFR2 by mutations R203C and W290C in the immunoglobulin-like (Ig-like) domains [41, 42].

Oncogenic mutations in isocitrate dehydrogenases

Mutations in isocitrate dehydogenases are also thought to contribute to the progression of cancer by altering the conformation of the protein. IDH1 and IDH2 catalyse the oxidative carboxylation of isocitrate to α-ketoglutarate. Mutational hotspots at R132H in IDH1, and R140Q and R172K in IDH2 alter the progression of this reaction. Recent structural work suggests that the R132H IDH1 mutation hampers the conformational change from the initial isocitrate binding state to the pre-transition state, thus causing an impairment of enzyme function [43]. This alters the progression of this reaction causing

the oncometabolite R(-)-2-hydroxyglutarate to be formed. R(-)-2-hydroxyglutarate is implicated in genomic hypermethylation, leading to histone methylation, genomic instability, and finally malignant transformation [44].

Domain-based approaches at identifying mutational hotspots

Although most of the analysis of cancer mutations is based around a gene centric view, a few studies have focused on domain-based analyses [45, 46][50] and they may be particularly fruitful when studying mechanisms of activation of proteins. Larger proteins comprise recognizable smaller sequence domains, which recur in other proteins in various combinations. These domains may be thought of as units of evolution, creating protein domain families, and have evolved from a common ancestor. As a domain can exist across multiple proteins with conserved function and structure, it follows that similarly located mutations across different proteins in the same domain should have similar effects on the function of that domain. Proteome-wide analyses have been performed to identify domains enriched in missense mutations [45, 47] [50] and to identify domain-centric positions of hotspot missense mutations [48, 49] [50]. These studies focused exclusively on missense mutation and as yet, little attempt was to use these data to distinguish between activating and loss of function mutations in the majority of cases.

We are currently mapping all simple small mutations (missense, truncations and indels) from over thirty different types of cancer to equivalent positions in multiple sequence alignments of protein domains. These data are being used to identify domain-centric mutational hotspots and can be accessed through

the MOKCa database.

Using the biological knowledge associated with protein domains, such as structural information and evolutionary conservation, will enable us to understand the functional consequences of infrequent mutations in well-characterised domain families and will facilitate additional insights into the roles of these mutations in cancer.

Figure Legends

Figure 1: This is an illustration of the data visualisation available on the different webpages on MOKCa web-interface. Figure (a) shows sets of cancer-related genes that can be browsed by gene name. Figure (b) shows a schematic diagram of the domain architecture of the protein (BRAF) with the positions of somatic mutations mapped to the protein sequence. Blue lines indicate missense mutations, dotted black lines indicate silent mutations and triangles are used to show insertions (pointing down) or deletions (pointing up). In frame indels are coloured blue, and frame shift indels are coloured green, solid black lines indicate nonsense mutations. Figure (c) is an extract from the summary table for mutation aggregates. As well as describing the mutations and their frequency it also indicates which domain the mutation is in, whether it is near any post-translational modifications and highlights which cancers it is found in. Figure (d) shows in more detail the post-translational modifications near the mutation. Figure (e) highlights the position of the mutation within a protein structure. In the example shown, the domain containing the mutation, a protein kinase domain (Pkinase), is coloured in red,

and the mutated residue is displayed as a space filling model. Figure (f) displays the distinct number of protein coding mutations (aggregates) found in each gene.

Figure 2: This is a schematic illustration of the change in the equilibrium of the active/open and inactive/closed conformational states of protein kinases. Figure (a) shows the default equilibrium of a protein kinase. When the activation loop is phosphorylated, the active conformer is stabilised and the equilibrium moves toward the active conformation. This is illustrated in figure (b). Activating mutations have a tendency to destabilise the inactive conformation also moving the equilibrium towards the active conformation. This is illustrated in figure (c).

Figure 3: Figures (a) and (b) show the structural impact of the V600E mutation in the protein product of BRAF as predicted by the SAAP [15] algorithm. The predicted impact of the mutation differs significantly dependent on whether the protein is in the (a) active or (b) inactive protein kinase conformation. Figures (c) and (d) show the predicted positions of the V600E mutation within the protein structure. The position of the mutated residue also differs significantly depending on whether the protein is in the (c) active or (d) inactive protein kinase conformation. Figure (c) is modelled on the PDB template 3PSD, chain B, and figure (d) on 3SKC chain B.

Key Words

Mutations, Oncogene, Tumour suppressor

Funding

FP was supported by a Daphne Jackson Fellowship funded by the MRC, GB-H is in receipt of an MRC studentship and HB holds a PhD studentship funded be the Saudi Arabian Ministry of Higher Education.

References

1. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res, 43, D789-798.
2. Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer. Cell, 100, 57-70.
3. Jeggo, P.A., Pearl, L.H., and Carr, A.M. (2016) DNA repair, genome stability and cancer: a historical perspective. Nat Rev Cancer, 16, 35-42.
4. Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009) The cancer genome. Nature, 458, 719-724.
5. Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. Cell, 144, 646-674.

6. Yap, T.A. and Workman, P. (2012) Exploiting the cancer genome: strategies for the discovery and clinical development of targeted molecular therapeutics. Annu Rev Pharmacol Toxicol, 52, 549-573.
7. Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C.Y., Jia, M., De, T., Teague, J.W., Stratton, M.R., McDermott, U., and Campbell, P.J. (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res, 43, D805-811.
8. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., Kiezun, A., Hammerman, P.S., McKenna, A., Drier, Y., Zou, L., Ramos, A.H., Pugh, T.J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortes, M.L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D.I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A.M., Lohr, J., Landau, D.A., Wu, C.J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S.A., Mora, J., Lee, R.S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S.B., Roberts, C.W., Biegel, J.A., Stegmaier, K., Bass, A.J., Garraway, L.A., Meyerson, M., Golub, T.R., Gordenin, D.A., Sunyaev, S., Lander, E.S., and Getz, G. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature, 499, 214-218.

9. Greenman, C., Wooster, R., Futreal, P.A., Stratton, M.R., and Easton, D.F. (2006) Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics, 173, 2187-2198.
10. Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., Yates, L.R., Papaemmanuil, E., Beare, D., Butler, A., Cheverton, A., Gamble, J., Hinton, J., Jia, M., Jayakumar, A., Jones, D., Latimer, C., Lau, K.W., McLaren, S., McBride, D.J., Menzies, A., Mudie, L., Raine, K., Rad, R., Chapman, M.S., Teague, J., Easton, D., Langerod, A., Oslo Breast Cancer, C., Lee, M.T., Shen, C.Y., Tee, B.T., Huimin, B.W., Broeks, A., Vargas, A.C., Turashvili, G., Martens, J., Fatima, A., Miron, P., Chin, S.F., Thomas, G., Boyault, S., Mariani, O., Lakhani, S.R., van de Vijver, M., van 't Veer, L., Foekens, J., Desmedt, C., Sotiriou, C., Tutt, A., Caldas, C., Reis-Filho, J.S., Aparicio, S.A., Salomon, A.V., Borresen-Dale, A.L., Richardson, A.L., Campbell, P.J., Futreal, P.A., and Stratton, M.R. (2012) The landscape of cancer genes and mutational processes in breast cancer. Nature, 486, 400-404.
11. Garraway, L.A. and Lander, E.S. (2013) Lessons from the cancer genome. Cell, 153, 17-37.
12. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004) A census of human cancer genes. Nat Rev Cancer, 4, 177-183.
13. Brauch, H., Pomer, S., Hieronymus, T., Schadt, T., Lohrke, H., and Komitowski, D. (1994) Genetic alterations in sporadic renal-cell

carcinoma: molecular analyses of tumor suppressor gene harboring chromosomal regions 3p, 5q, and 17p. World J Urol, 12, 162-168.

14. Wan, P.T., Garnett, M.J., Roe, S.M., Lee, S., Niculescu-Duvaz, D., Good, V.M., Jones, C.M., Marshall, C.J., Springer, C.J., Barford, D., Marais, R., and Cancer Genome, P. (2004) Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. Cell, 116, 855-867.
15. Al-Numair, N.S. and Martin, A.C. (2013) The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. BMC Genomics, 14 Suppl 3, S4.
16. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013) Cancer genome landscapes. Science, 339, 1546-1558.
17. Ng, P.C. and Henikoff, S. (2001) Predicting deleterious amino acid substitutions. Genome Res, 11, 863-874.
18. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013) Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet, Chapter 7, Unit7 20.
19. Pires, D.E., Ascher, D.B., and Blundell, T.L. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res, 42, W314-319.
20. Yates, C.M., Filippis, I., Kelley, L.A., and Sternberg, M.J. (2014) SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol, 426, 2692-2701.

21. Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N., and Gaunt, T.R. (2013) Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics, 29, 1504-1510.
22. Reva, B., Antipin, Y., and Sander, C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res, 39, e118.
23. Gonzalez-Perez, A., Deu-Pons, J., and Lopez-Bigas, N. (2012) Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. Genome Med, 4, 89.
24. Espinosa, O., Mitsopoulos, K., Hakas, J., Pearl, F., and Zvelebil, M. (2014) Deriving a mutation index of carcinogenicity using protein structure and protein interfaces. PLoS One, 9, e84598.
25. Douville, C., Carter, H., Kim, R., Niknafs, N., Diekhans, M., Stenson, P.D., Cooper, D.N., Ryan, M., and Karchin, R. (2013) CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics, 29, 647-648.
26. Douville, C., Masica, D.L., Stenson, P.D., Cooper, D.N., Gygax, D.M., Kim, R., Ryan, M., and Karchin, R. (2015) Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). Hum Mutat.
27. Schroeder, M.P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2014) OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. Bioinformatics, 30, i549-555.

28. Richardson, C.J., Gao, Q., Mitsopoulous, C., Zvelebil, M., Pearl, L.H., and Pearl, F.M. (2009) MoKCa database--mutations of kinases in cancer. Nucleic Acids Res, 37, D824-831.
29. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A.J., Poux, S., Bougueleret, L., and Xenarios, I. (2016) UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. Methods Mol Biol, 1374, 23-54.
30. Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., and Bateman, A. (2015) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res.
31. Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res, 43, D512-520.
32. Gene Ontology, C. (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res, 43, D1049-1056.
33. Berman, H.M., Kleywegt, G.J., Nakamura, H., and Markley, J.L. (2012) The Protein Data Bank at 40: reflecting on the past to prepare for the future. Structure, 20, 391-396.
34. Altschul, S.F., Gertz, E.M., Agarwala, R., Schaffer, A.A., and Yu, Y.K. (2009) PSI-BLAST pseudocounts and the minimum description length principle. Nucleic Acids Res, 37, 815-824.

35. McMahon, B. and Hanson, R.M. (2008) A toolkit for publishing enhanced figures. J Appl Crystallogr, 41, 811-814.
36. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics, 25, 1189-1191.
37. Pearl, L.H., Schierz, A.C., Ward, S.E., Al-Lazikani, B., and Pearl, F.M. (2015) Therapeutic opportunities within the DNA damage response. Nat Rev Cancer, 15, 166-180.
38. Mitsopoulos, C., Schierz, A.C., Workman, P., and Al-Lazikani, B. (2015) Distinctive Behaviors of Druggable Proteins in Cellular Networks. PLoS Comput Biol, 11, e1004597.
39. Marino, K.A., Sutto, L., and Gervasio, F.L. (2015) The effect of a widespread cancer-causing mutation on the inactive to active dynamics of the B-Raf kinase. J Am Chem Soc, 137, 5280-5283.
40. Dhillon, A.S., Meikle, S., Peyssonnaux, C., Grindlay, J., Kaiser, C., Steen, H., Shaw, P.E., Mischak, H., Eychene, A., and Kolch, W. (2003) A Raf-1 mutant that dissociates MEK/extracellular signal-regulated kinase activation from malignant transformation and differentiation but not proliferation. Mol Cell Biol, 23, 1983-1993.
41. Reintjes, N., Li, Y., Becker, A., Rohmann, E., Schmutzler, R., and Wollnik, B. (2013) Activating somatic FGFR2 mutations in breast cancer. PLoS One, 8, e60264.
42. Lajeunie, E., Heuertz, S., El Ghouzzi, V., Martinovic, J., Renier, D., Le Merrer, M., and Bonaventure, J. (2006) Mutation screening in patients with syndromic craniosynostoses indicates that a limited number of

recurrent FGFR2 mutations accounts for severe forms of Pfeiffer syndrome. Eur J Hum Genet, 14, 289-298.

43. Yang, B., Zhong, C., Peng, Y., Lai, Z., and Ding, J. (2010) Molecular mechanisms of "off-on switch" of activities of human IDH1 by tumor-associated mutation R132H. Cell Res, 20, 1188-1200.
44. Kato, Y. (2015) Specific monoclonal antibodies against IDH1/2 mutations as diagnostic tools for gliomas. Brain Tumor Pathol, 32, 3-11.
45. Nehrt, N.L., Peterson, T.A., Park, D., and Kann, M.G. (2012) Domain landscapes of somatic mutations in cancer. BMC Genomics, 13 Suppl 4, S9.
46. Porta-Pardo, E. and Godzik, A. (2014) e-Driver: a novel method to identify protein regions driving cancer. Bioinformatics, 30, 3109-3114.
47. Peterson, T.A., Nehrt, N.L., Park, D., and Kann, M.G. (2012) Incorporating molecular and functional context into the analysis and prioritization of human variants associated with cancer. J Am Med Inform Assoc, 19, 275-283.
48. Peterson, T.A., Adadey, A., Santana-Cruz, I., Sun, Y., Winder, A., and Kann, M.G. (2010) DMDM: domain mapping of disease mutations. Bioinformatics, 26, 2458-2459.
49. Yue, P., Forrest, W.F., Kaminker, J.S., Lohr, S., Zhang, Z., and Cavet, G. (2010) Inferring the functional effects of mutation through clusters of mutations in homologous proteins. Hum Mutat, 31, 264-271.

50 Miller, M.L., Reznik, E., Gauthier, N.P., Aksoy, B.A., Korkut, A., Gao, J., Cirello, G., Schultz, N. and Sander, C. (2015) Pan-cancer analysis of mutation hotspots in protein domains. Cell Systems 1,197-209.

(a)

BRAF: P15056

(b)

600	p.V600E	Missense	19548	Pkinase_Tyr	2 P , 1 U	PDB	Agl CNS Eye Pty Adr Thy Bil GIT LIn Oes Pnc Sln Sto UAT Bon Ski SoT Bre GTr Ova Pro Tes HLT Lvr Lng UTr NS
601	p.K601I	Missense	1	Pkinase_Tyr	2 P , 1 U	PDB	Ski
601	p.K601R	Missense	1	Pkinase_Tyr	2 P , 1 U	PDB	Ski

(c)

Mutation : *P15056* p.V600E

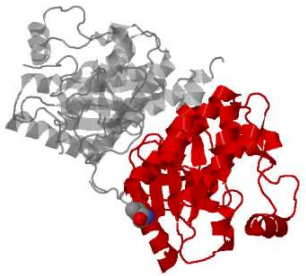
Mutation type : Missense

Domain : *PF07714* Pkinase_Tyr

Residue modifications at or near mutation

RESIDUE NUMBER	RESIDUE	MODIFICATION
599	T	phosphorylation
601	K	ubiquitination
602	S	phosphorylation

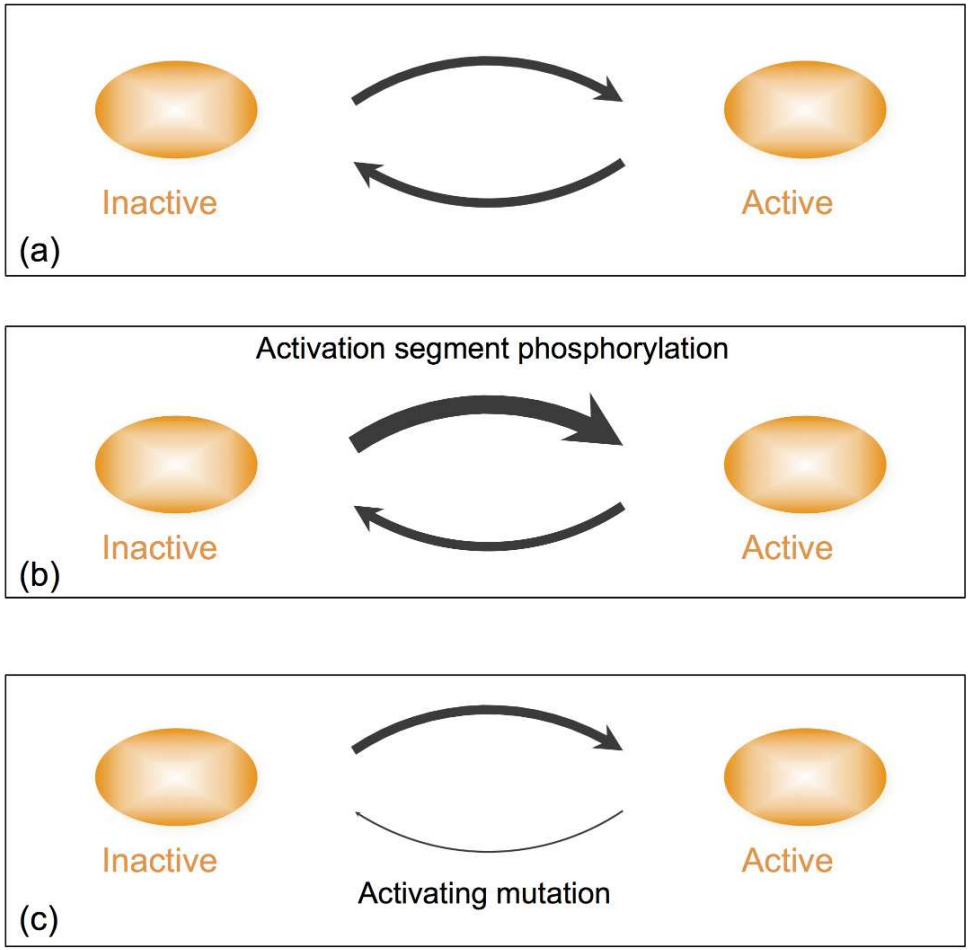
(d)



(e)

COSMIC GENE NAME	NUMBER OF DISTINCT MUTATIONS
BCKDK	2
BCR	13
BLK	5
BMPR1A	10
BMPR1B	14
BMPR2	16
BMX	6
BRAF	157

(f)

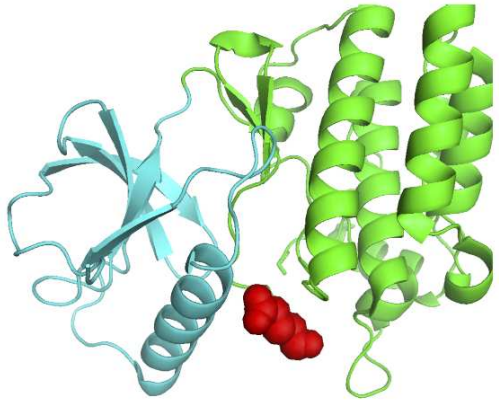

Inactive
Active
(a)
Activation segment phosphorylation
Inactive
Active
(b)
Inactive
Active
Activating mutation
(c)

3psd Residue B600 | Structure type: crystal | Resolution: 3.60A | R-factor: 28.00% ⚠	
HBonds	Binding
No problems identified	No problems identified
BuriedCharge	Voids
No problems identified	No problems identified
SProtFT	SurfacePhobic
No problems identified	No problems identified
Interface	Glycine
No problems identified	No problems identified
Clash	CisPro
No problems identified	No problems identified
Proline	CorePhilic
No problems identified	No problems identified
Impact	SSGeom
No problems identified	No problems identified.

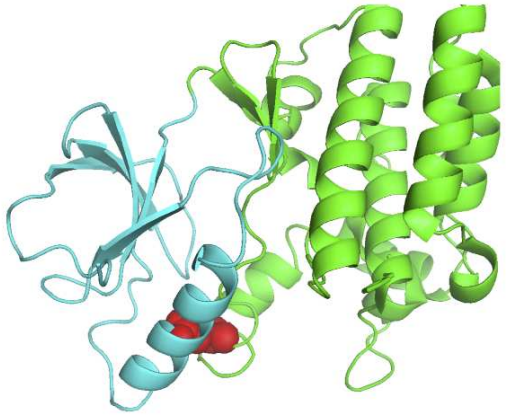
(a)

3skc Residue B600 | Structure type: crystal | Resolution: 3.20A | R-factor: 23.50% ⚠	
HBonds	Binding
No problems identified	No problems identified
BuriedCharge	Voids
The mutation resulted in introducing or removing a buried charge.	No problems identified
SProtFT	SurfacePhobic
No problems identified	No problems identified
Interface	Glycine
No problems identified	No problems identified
Clash	CisPro
No problems identified	No problems identified
Proline	CorePhilic
No problems identified	The mutation introduces a hydrophilic residue into the core of the protein.
Impact	SSGeom
No problems identified	No problems identified.

(b)



(c)



(d)

Supplementary Figure 1

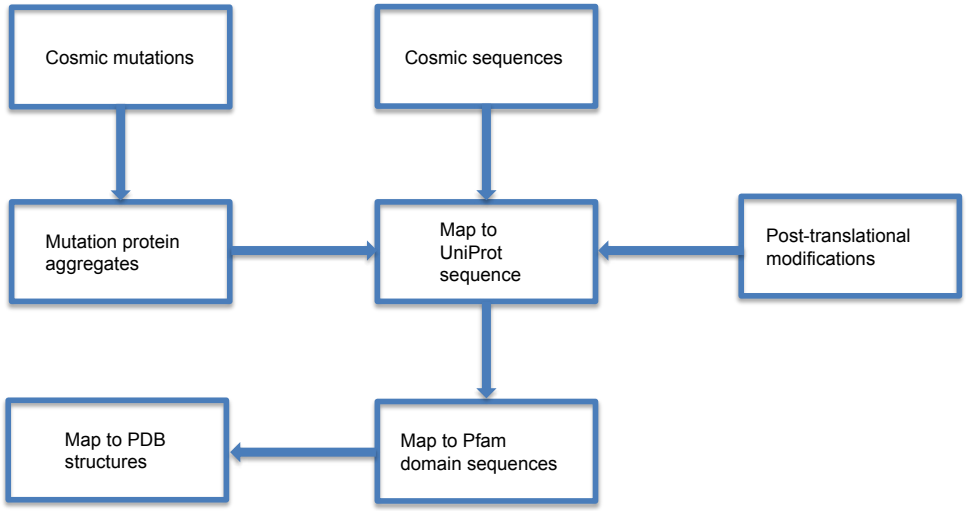
This figure outlines the steps required to populate the MoKCA database.

Mutation mapping: All Cosmic mutations are analysed at the protein level and clustered into aggregate mutations. The positions of these mutations are then re-mapped onto the UniProt protein sequence using a Cosmic to UniProt pairwise protein sequence alignment.

Sequence Alignments: Protein sequences downloaded from the Cosmic database are scanned against all human UniProt sequences. A pairwise sequence alignment is obtained for each Cosmic sequence to the nearest UniProt sequence found.

Pfam domain assignments: Domain boundaries for UniProt sequences are extracted from the Pfam database and domain sequence files constructed. Each domain sequence is then scanned against the PDB sequence library and the best ten matches are then realigned using a dynamic programming algorithm. These domain sequence alignments are used to map both the Pfam and mutational data onto the PDB structures for visualisation on the web-pages.

Posttranslational modifications are directly mapped onto UniProt protein sequences. Other functional annotation is extracted from external databases using the UniProt accession code.


Cosmic mutations
Cosmic sequences
Mutation protein aggregates
Map to UniProt sequence
Post-translational modifications
Map to PDB structures
Map to Pfam domain sequences